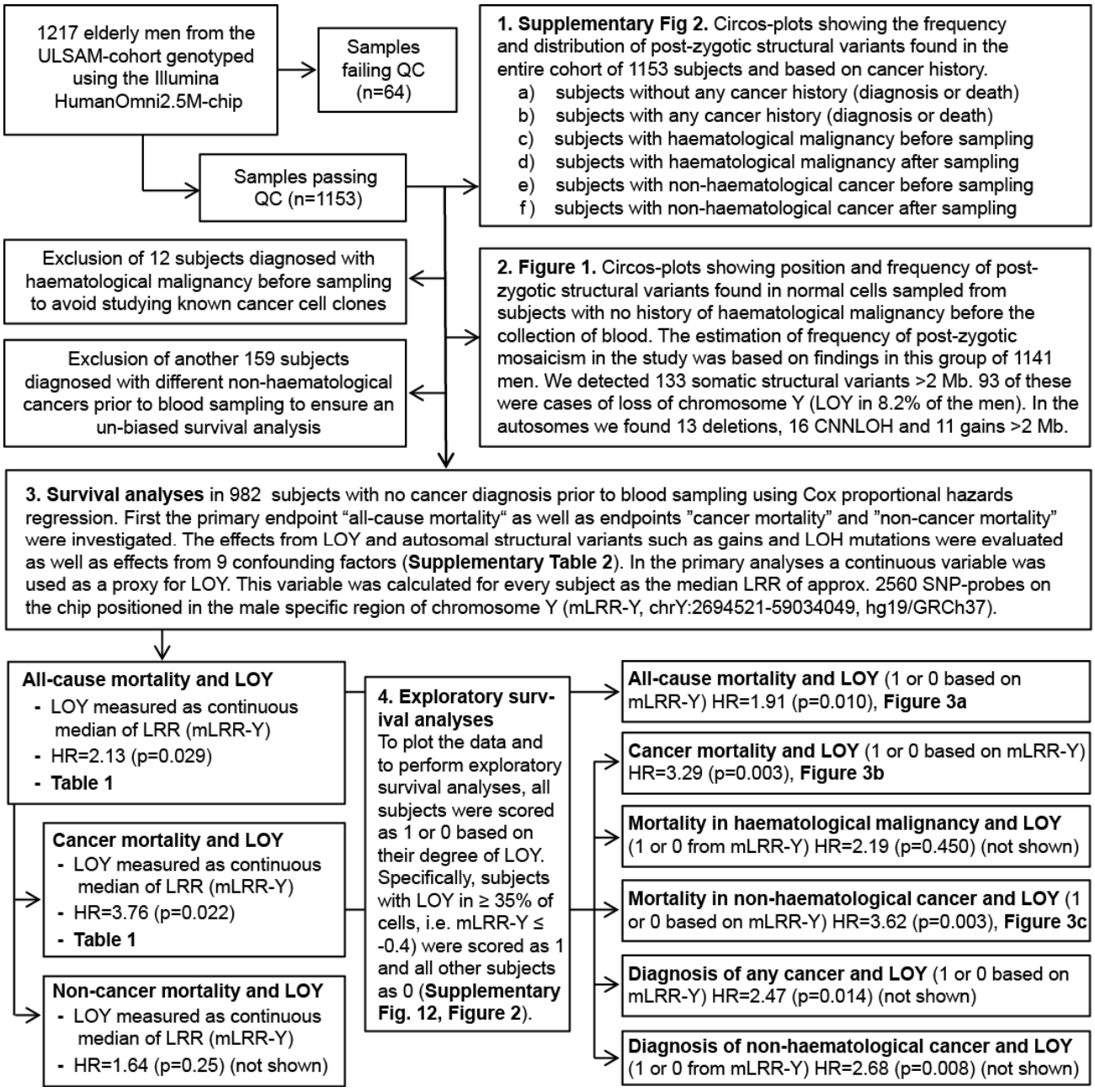
Supplementary Information

Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer

Lars A. Forsberg, Chiara Rasi, Niklas Malmqvist, Hanna Davies, Saichand Pasupulati, Geeta Pakalapati, Johanna Sandgren, Teresita Diaz de Ståhl, Ammar Zaghlool, Vilmantas Giedraitis, Lars Lannfelt, Joannah Score, Nicholas C.P. Cross, Devin Absher, Eva Tiensuu Janson, Cecilia M. Lindgren, Andrew P. Morris, Erik Ingelsson, Lars Lind, and Jan P. Dumanski

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 1

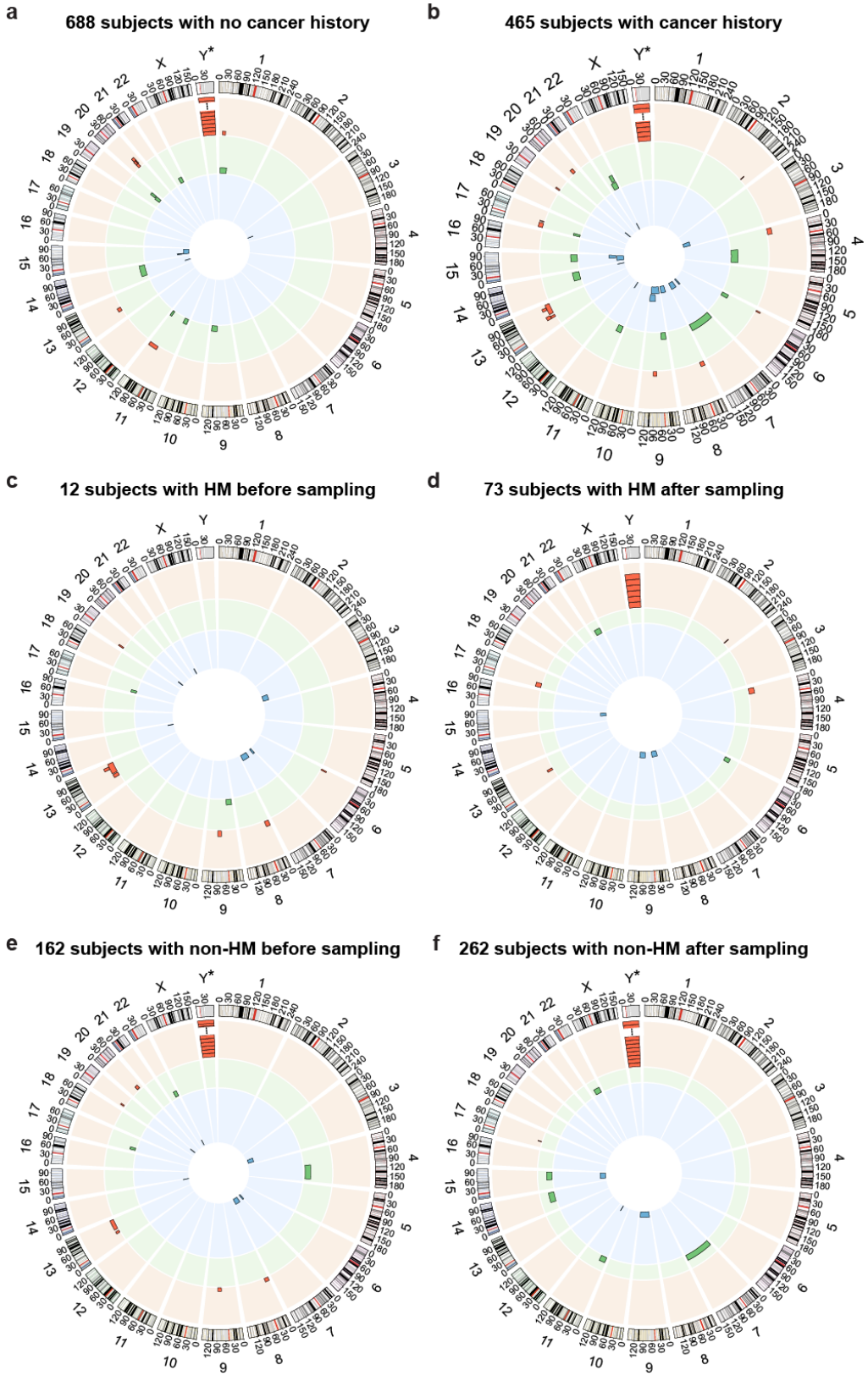


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 1. Flow-chart describing all steps in analysis of 1217 ULSAM participants genotyped on Illumina's 2.5MHumanOmni SNP-beadchip. This figure also summarizes major findings and refers to relevant Figures and Tables showing detailed results. Boxes with numbers 1 and 2 summarize scoring of aberrations from the cohorts of 1153 and 1141 ULSAM subjects. Boxes with numbers 3 and 4 refer to primary survival analyses, respectively, exploratory survival analyses using 982 participants.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 2

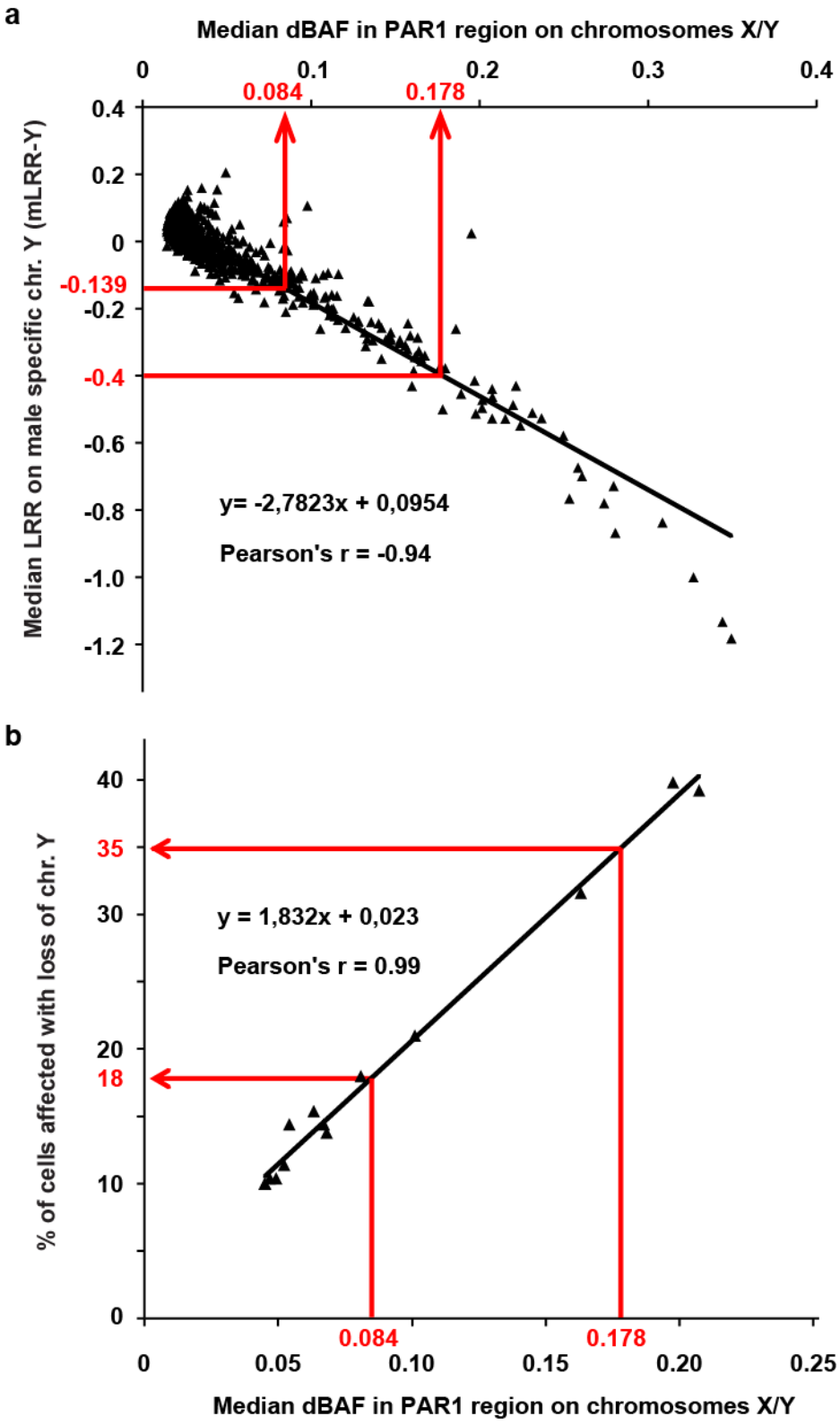


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 2. Circos-plots showing the structural variants found in the entire cohort of 1153 participants that were successfully genotyped on Illumina beadchips. The number of deletions, CNNLOH and gains are shown with red, green and blue bars, respectively. Panels a and b show 688 subjects without cancer history and 465 cases with cancer diagnoses, respectively. Panels c and d display 12 participants with history of haematological malignancy (HM) before blood sampling and 73 individuals with diagnoses of haematological malignancy after blood sampling. Correspondingly, panels e and f illustrate data from 162 participants with non-haematological (non-HM) malignancy diagnoses prior to blood sampling and 262 cases who received diagnoses of non-haematological cancer after sampling. Data showing LOY in panels a, b, e and f are not shown to scale (highlighted with an asterisk; *). The numbers for LOY events in these panels are 55, 38, 13, and 29, respectively.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 3

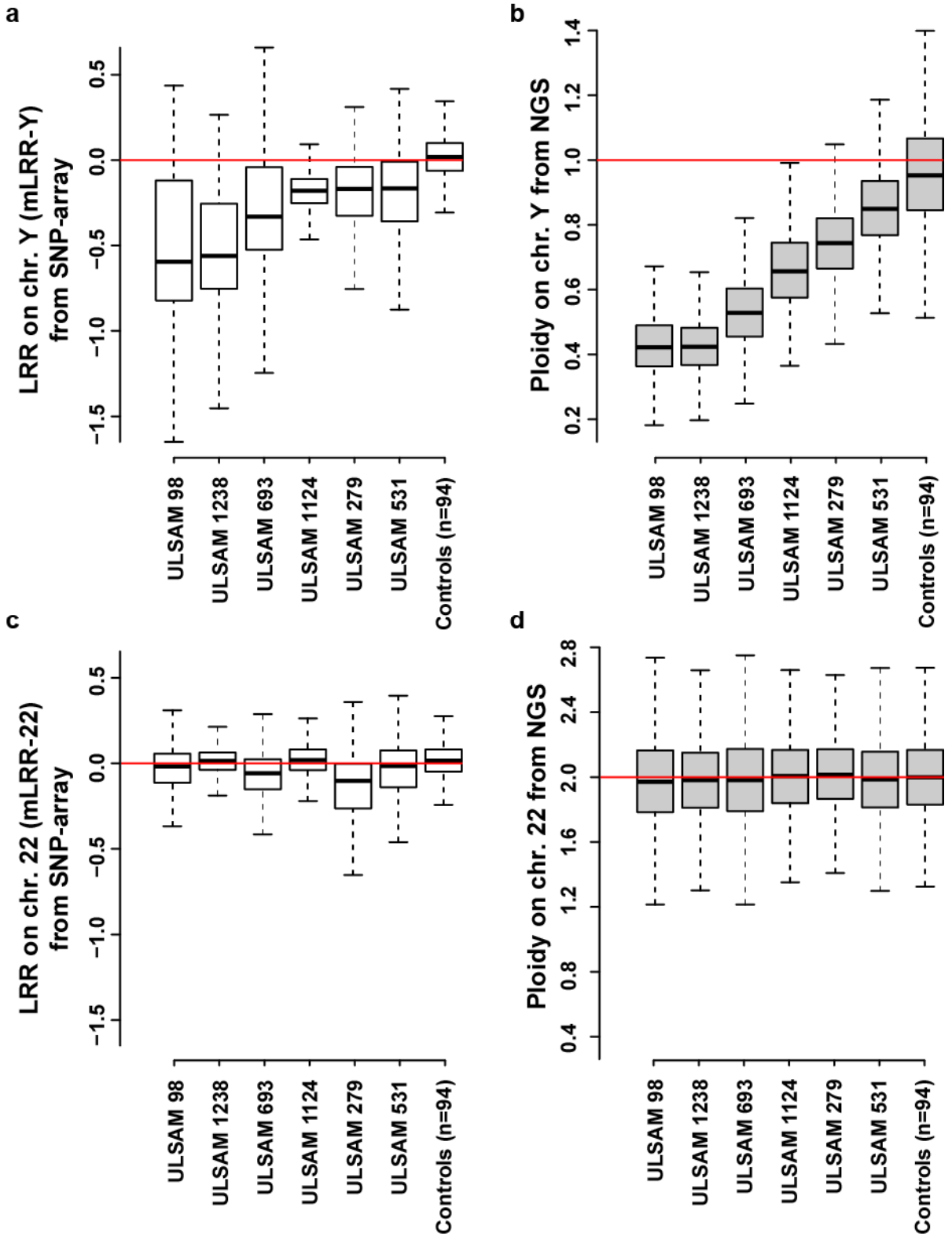


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 3. Estimation of the percentage of blood cells affected with loss of chromosome Y (LOY) through analysis of SNP-array data from the pseudoautosomal region 1 (PAR1) of chromosomes X/Y using MAD-software[35] in the ULSAM cohort. PAR1 is the largest of the PARs (regions with homologous sequences on chromosomes X and Y) with coordinates 10001-2649520 on Y and 60001-2699520 on X. MAD-software is a tool for detection and quantification of somatic structural variants from SNP-array data, which uses diploid B-allele frequency (BAF) for identification and Log R Ratio (LRR) for quantification of somatic variants and is not originally intended for analyses of chromosome Y data. However, by using the correlation between the LRR in the PAR1-region of Y and the dBAF (i.e. the absolute deviation from the expected BAF-value of 0.5 in heterozygous probes) of the PAR1-region of X/Y (panel a), we could use the MAD-quantification of the diploid PAR1 region on chromosomes X/Y to calculate the percentage of cells affected with LOY (panel b) in a two-step process. For example, the dBAF-value at the LRR-threshold for survival analyses (mLRR-Y $\leq$ -0.4) can be found using the equation given in panel a (i.e. 0.178). This equation ($y = -2.7823x + 0.0954$) is describing the relationship between mLRR-Y on Y and dBAF on XY for the 1141 subjects. Next, the percentage of cells affected by LOY can be found by applying the equation in panel b that describes the relationship between dBAF and the percentage of cells as estimated by the MAD software for 14 cases ($y = 1.832x + 0.023$). For this example, the dBAF of 0.178 translates to LOY in 35% of cells.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 4

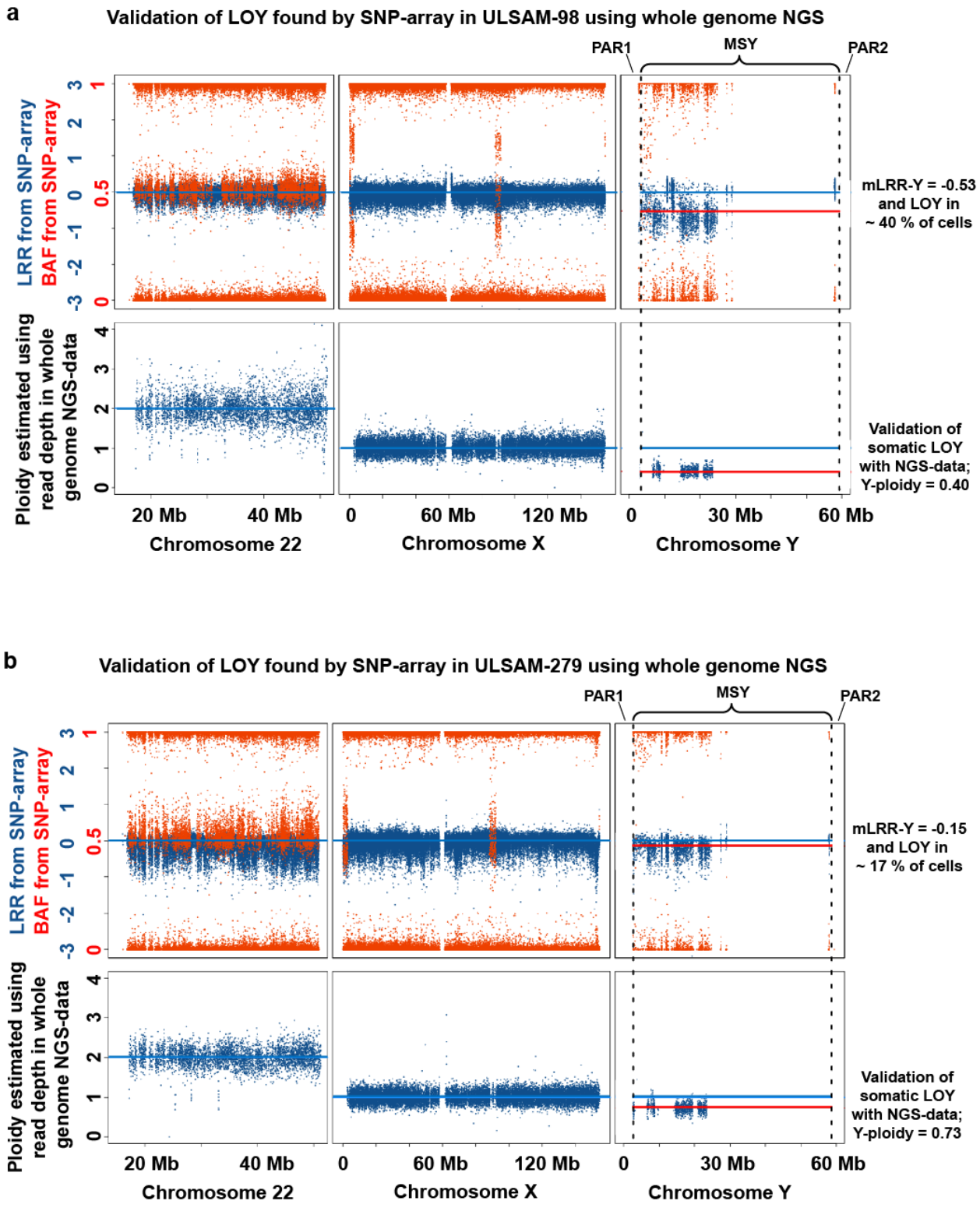


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 4. Validations of findings of LOY using next generation sequencing (NGS) for six candidate subjects. Low coverage whole genome sequencing was performed on 100 participants from the cohort. Among the 93 subjects with a median LRR in the male specific region on chromosome Y (mLRR-Y, i.e. the median Log R Ratio for ~2560 SNP-probes in the region chrY:2694521-59034049, hg19/GRCh37) lower than -0.139 (i.e. threshold for frequency estimation, **Fig. 2**), whole genome sequencing was performed in 6 participants. Panel a shows the Log R Ratio (LRR) data from the male specific region on chromosome Y (MSY) in these 6 subjects using boxplots. The rightmost box (in all panels) contains the data from the 94 sequenced individuals with an mLRR-Y above the -0.139 threshold for the frequency estimation. The red lines in all panels represent the expected normal state. The NGS-data from the 6 subjects and the 94 controls are plotted in panel b. The median read-depth in the MSY of the 94 subjects without LOY was 1.6 (standard deviation (SD)=0.6). The corresponding read-depth in the 6 subjects with LOY was 1.3 (SD=0.5). In comparison, the median read-depth on chromosome 22 was 3.8 (SD=1.4) in the 94 subjects without LOY and 3.8 (SD=1.2) in the 6 subjects with LOY. The read-depth data was used to estimate the ploidy of chromosome 22 and the MSY-region on chromosome Y in comparison with the rest of the genome using the FREEC software[39]. The estimated ploidy is plotted in panels b and d. FREEC calculates ploidy for the regions of interest as the copy number value in each of 5 kb windows in the region of interest after GC-content read count normalization, given a normal autosomal ploidy of 2. Panel c and d show that the copy number state on chromosome 22 is normal in the participants affected with LOY and plotted in panels a and b, using SNP-array and NGS-data, respectively.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 5

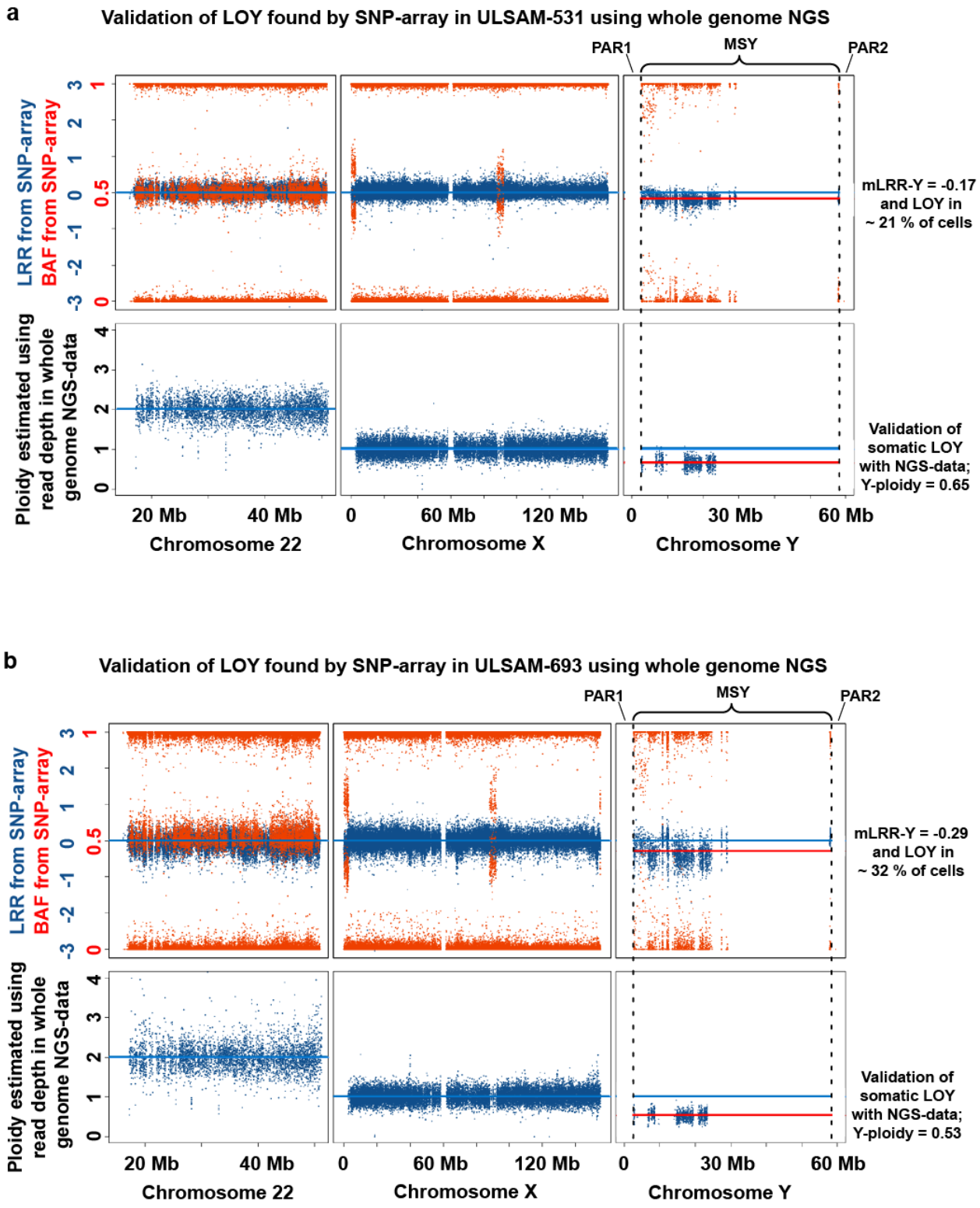


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 5. Detailed individual validations of LOY in ULSAM subjects 98 and 279 using low coverage whole genome next generation sequencing (NGS). In panels a and b are plotted the SNP-array and NGS data from chromosomes 22, X and Y for each subject. The LRR (blue dots) and B-Allele Frequency (BAF, red dots) from the SNP-array are plotted overlaid and the percentages of cells affected were calculated using MAD-software[35]. LRR values on sex chromosomes were normalized to a diploid state and chromosome X probes (residing outside PAR regions) with ambiguous clustering (scored as heterozygotes) are excluded from analyses. The ploidy-value, estimated from the NGS-data, was calculated using FREEC-software[39]. Blue lines indicate the normal copy number state and red line the observed LOY. SNP-array data within MSY include clusters of probes for known highly repetitive genes/loci, e.g. the *TSPYA*, *TSPYB* and *RBMY1* genes[16,17]. Therefore, probes covering these loci do not reflect true copy-number state of chromosome Y. These regions are not included in mapping of the NGS sequence reads.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 6

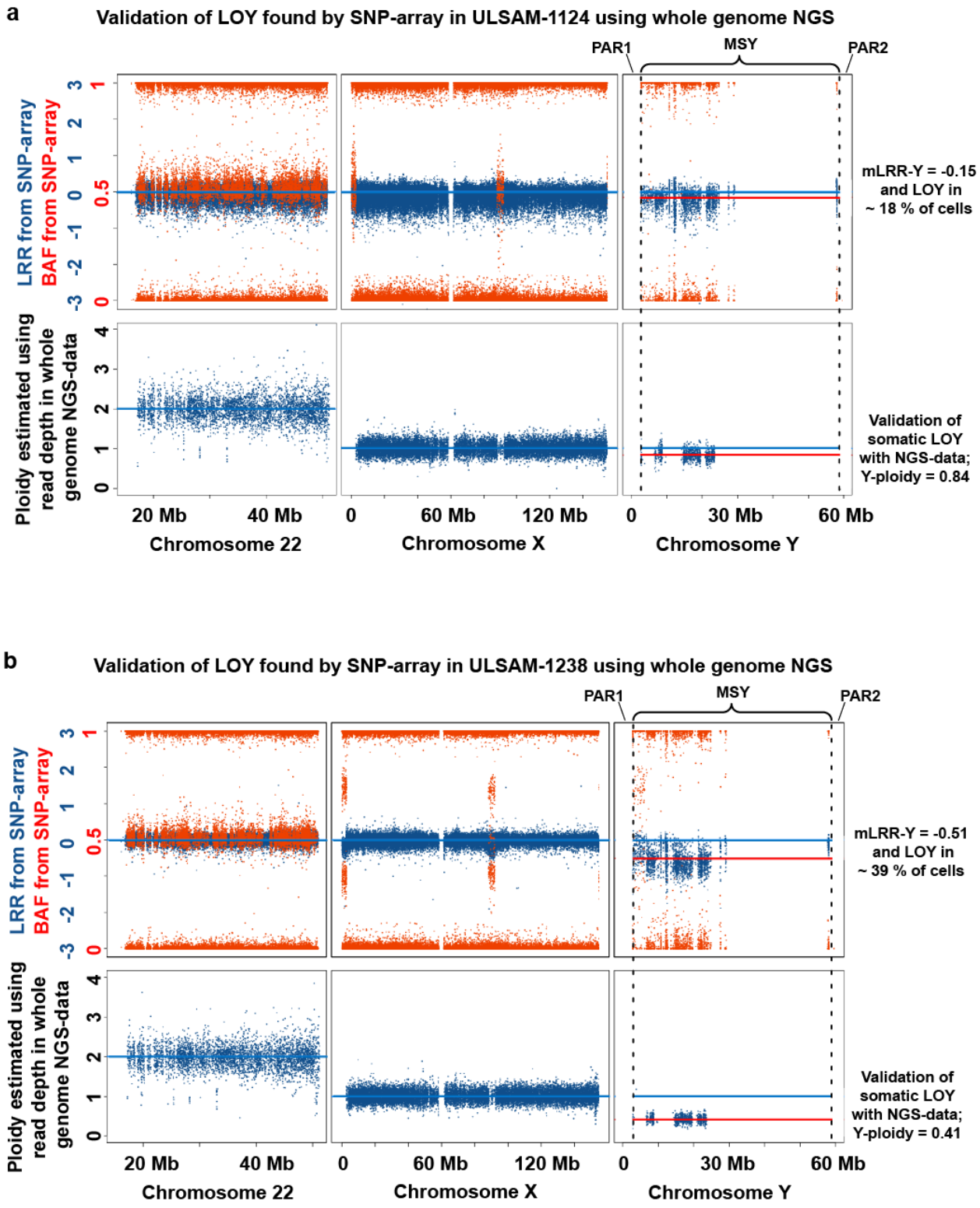


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 6. Detailed individual validations of LOY in ULSAM subjects 531 and 693 using low coverage whole genome next generation sequencing (NGS). In panels a and b are plotted the SNP-array and NGS data from chromosomes 22, X and Y for each subject. The LRR (blue dots) and B-Allele Frequency (BAF, red dots) from the SNP-array are plotted overlaid and the percentages of cells affected were calculated using MAD-software[35]. LRR values on sex chromosomes were normalized to a diploid state and chromosome X probes (residing outside PAR regions) with ambiguous clustering (scored as heterozygotes) are excluded from analyses. The ploidy-value, estimated from the NGS-data, was calculated using FREEC-software[39]. Blue lines indicate the normal copy number state and red line the observed LOY. SNP-array data within MSY include clusters of probes for known highly repetitive genes/loci, e.g. the *TSPYA*, *TSPYB* and *RBMY1* genes[16,17]. Therefore, probes covering these loci do not reflect true copy-number state of chromosome Y. These regions are not included in mapping of the NGS sequence reads.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 7

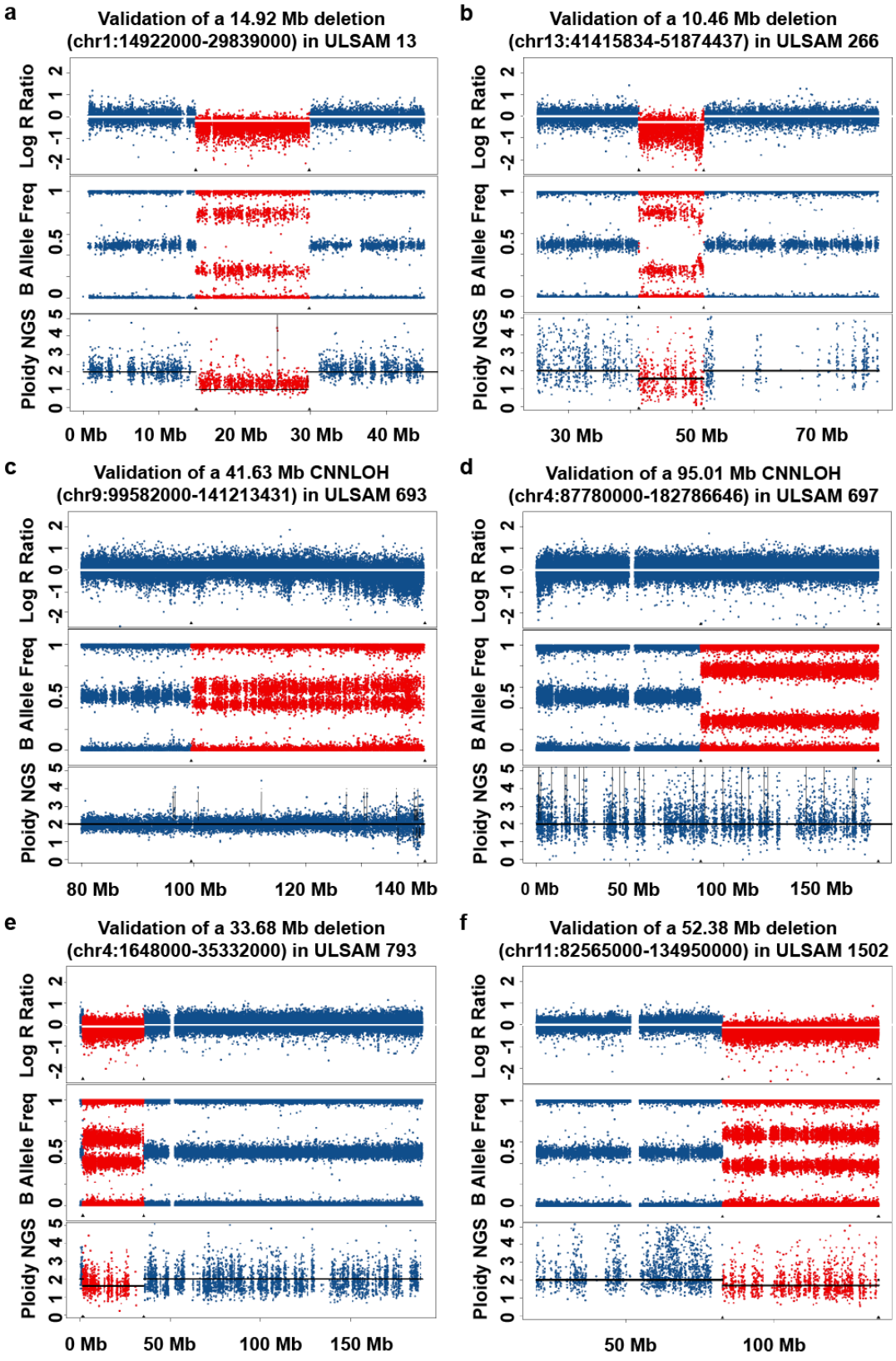


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 7. Detailed individual validations of LOY in ULSAM subjects 1124 and 1238 using low coverage whole genome next generation sequencing (NGS). In panels a and b are plotted the SNP-array and NGS data from chromosomes 22, X and Y for each subject. The LRR (blue dots) and B-Allele Frequency (BAF, red dots) from the SNP-array are plotted overlaid and the percentages of cells affected were calculated using MAD-software[35]. LRR values on sex chromosomes were normalized to a diploid state and chromosome X probes (residing outside PAR regions) with ambiguous clustering (scored as heterozygotes) are excluded from analyses. The ploidy-value, estimated from the NGS-data, was calculated using FREEC-software[39]. Blue lines indicate the normal copy number state and red line the observed LOY. SNP-array data within MSY include clusters of probes for known highly repetitive genes/loci, e.g. the *TSPYA*, *TSPYB* and *RBMY1* genes[16,17]. Therefore, probes covering these loci do not reflect true copy-number state of chromosome Y. These regions are not included in mapping of the NGS sequence reads.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 8

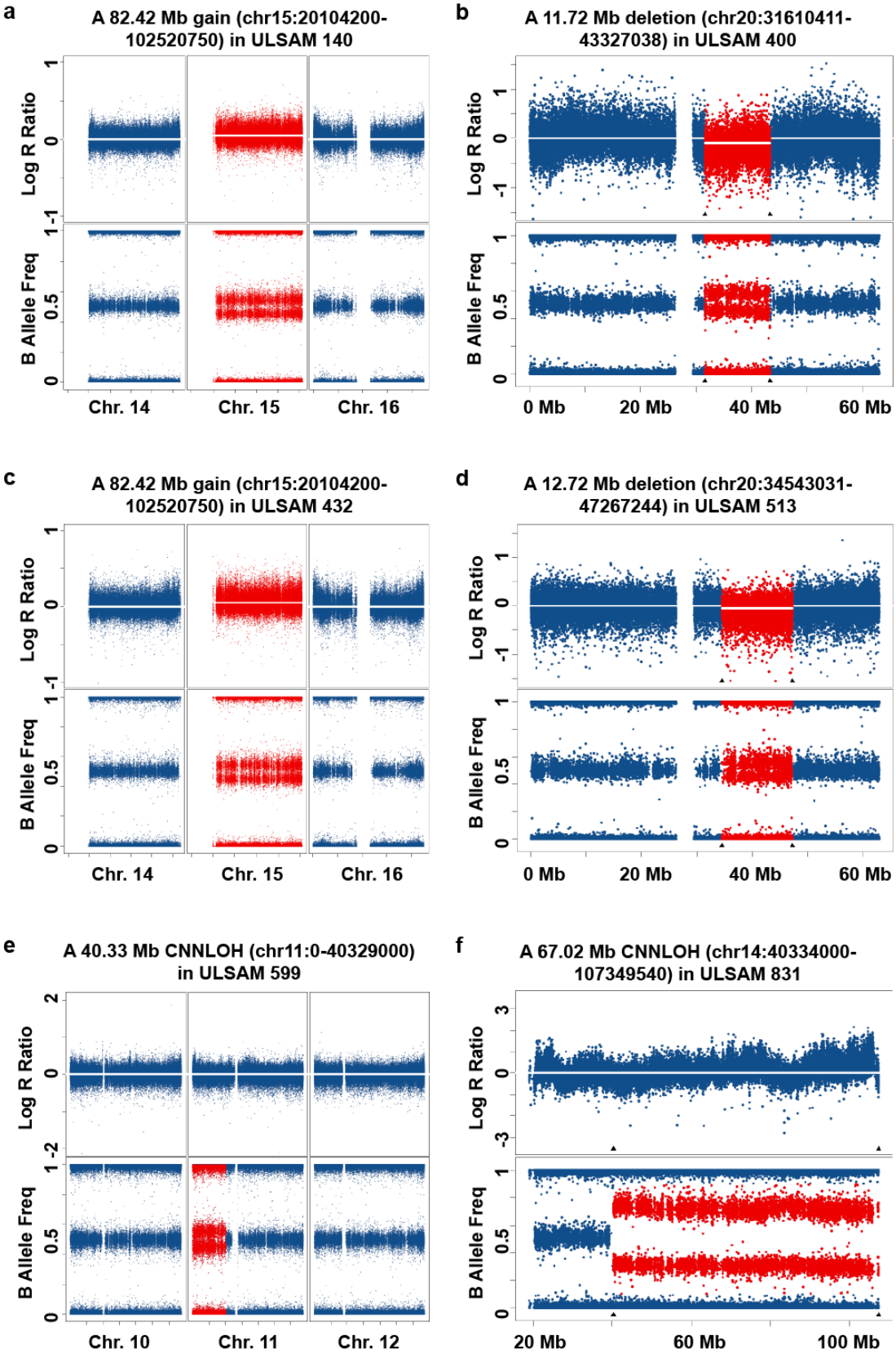


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 8. Examples of autosomal structural genetic aberrations ≥2 Mb in the ULSAM cohort. The Log R ratio (LRR) and the B allele frequency (BAF) from SNP-array as well as validations with next generation sequencing (NGS) using low coverage whole genome NGS (panel c) or exome sequencing (panels a, b, d-f) with an average coverage of 17x. Triangles indicate the positions of calls for structural variants and probes within these positions are plotted in red. Panels c and d show two examples of CNNLOH; in these cases NGS did not, as expected, detect any structural changes in these regions.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 9

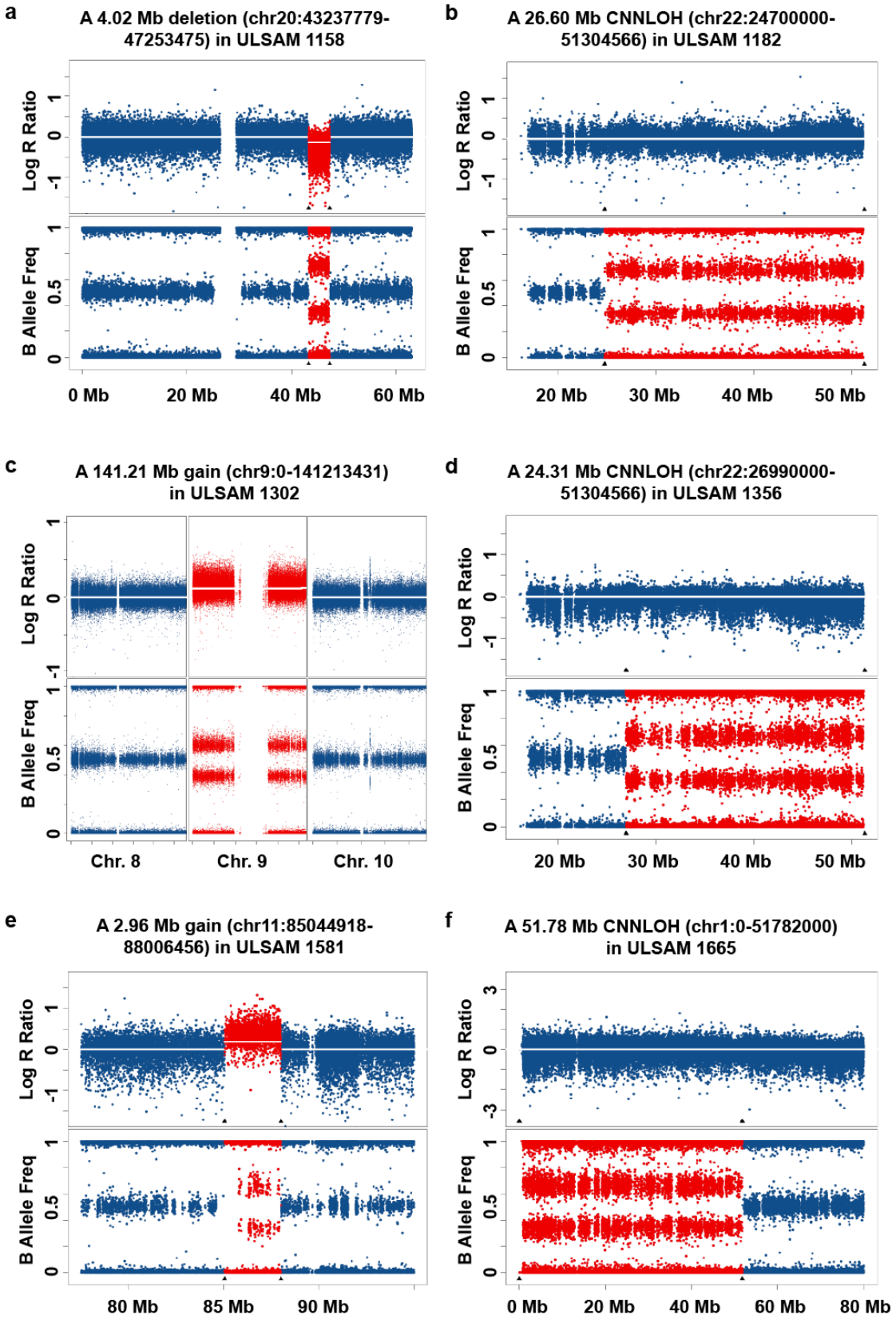


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 9. Examples of autosomal structural genetic aberrations ≥2 Mb in the ULSAM cohort. In each panel the Log R ratio (LRR) and the B allele frequency (BAF) from SNP-array are plotted. Triangles indicate the positions of calls for structural variants and probes within these positions are plotted in red.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 10

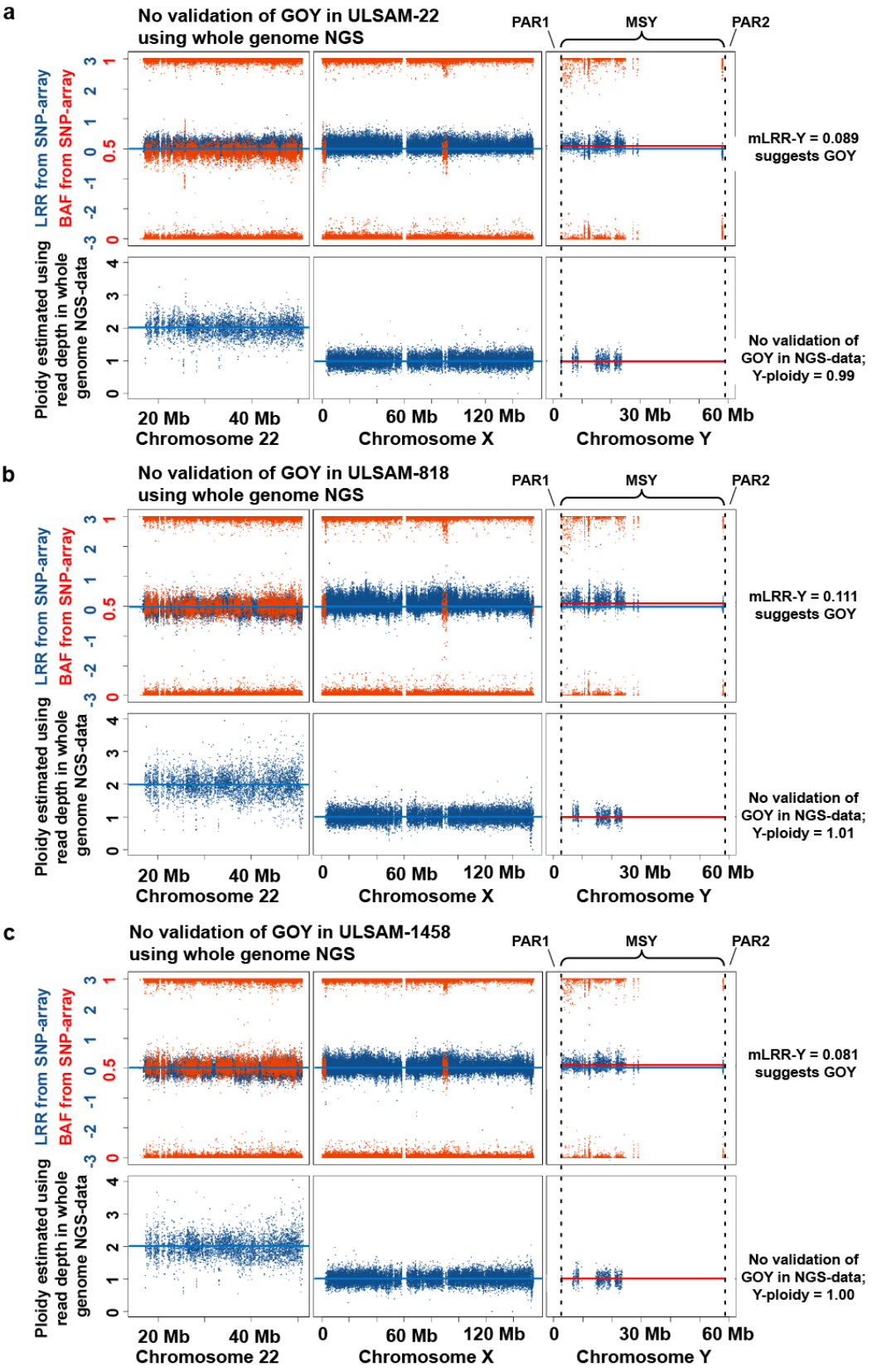


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 10. Examples of autosomal structural genetic aberrations ≥2 Mb in the ULSAM cohort. In each panel the Log R ratio (LRR) and the B allele frequency (BAF) from SNP-array are plotted. Triangles indicate the positions of calls for structural variants and probes within these positions are plotted in red.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 11

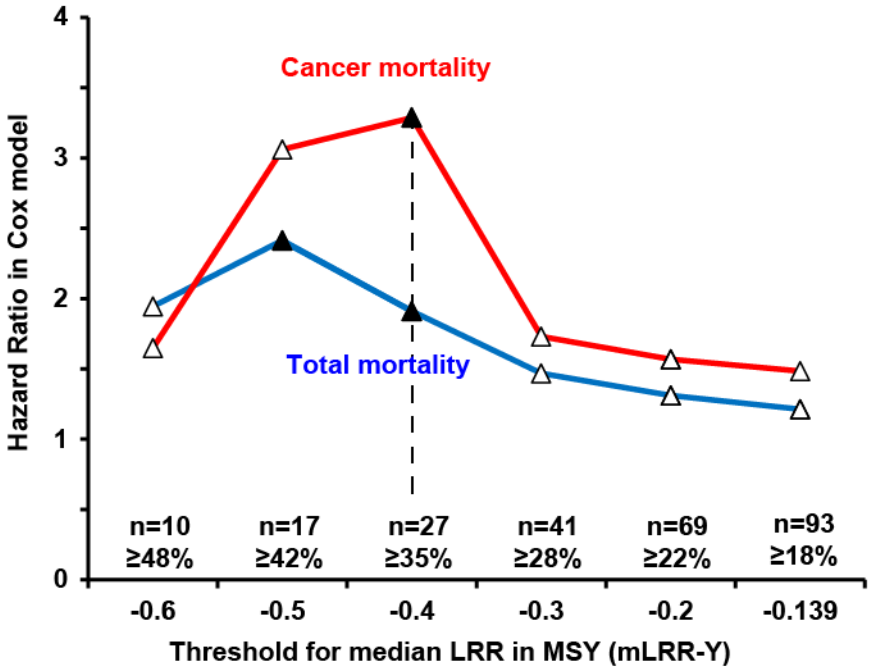


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 11. No validation of suggested cases of gain of chromosome Y (GOY) using low coverage (~5x) whole genome next generation sequencing (NGS). Of 100 sequenced participants, 3 had a positive median Log R Ratio (LRR) on the SNP-array in the male specific part of chromosome Y (mLRR-Y) indicating a possible gain of chromosome Y. In panel a-c are plotted the SNP-array and NGS data from chromosomes 22, X and Y for each of these three subjects. The LRR and B-Allele Frequency (BAF) from SNP-array are plotted overlaid and the percentages of cells affected were calculated using MAD-software[35]. LRR on sex chromosomes were normalized to diploid state and chromosome X probes (residing outside PAR regions) with ambiguous clustering (scored as heterozygotes) are excluded from analyses. The ploidy estimated from the NGS-data was calculated using FREEC-software[39]. Blue lines indicate the normal copy number state and red line the observed LOY.

Nature Genetics: doi:10.1038/ng.2966

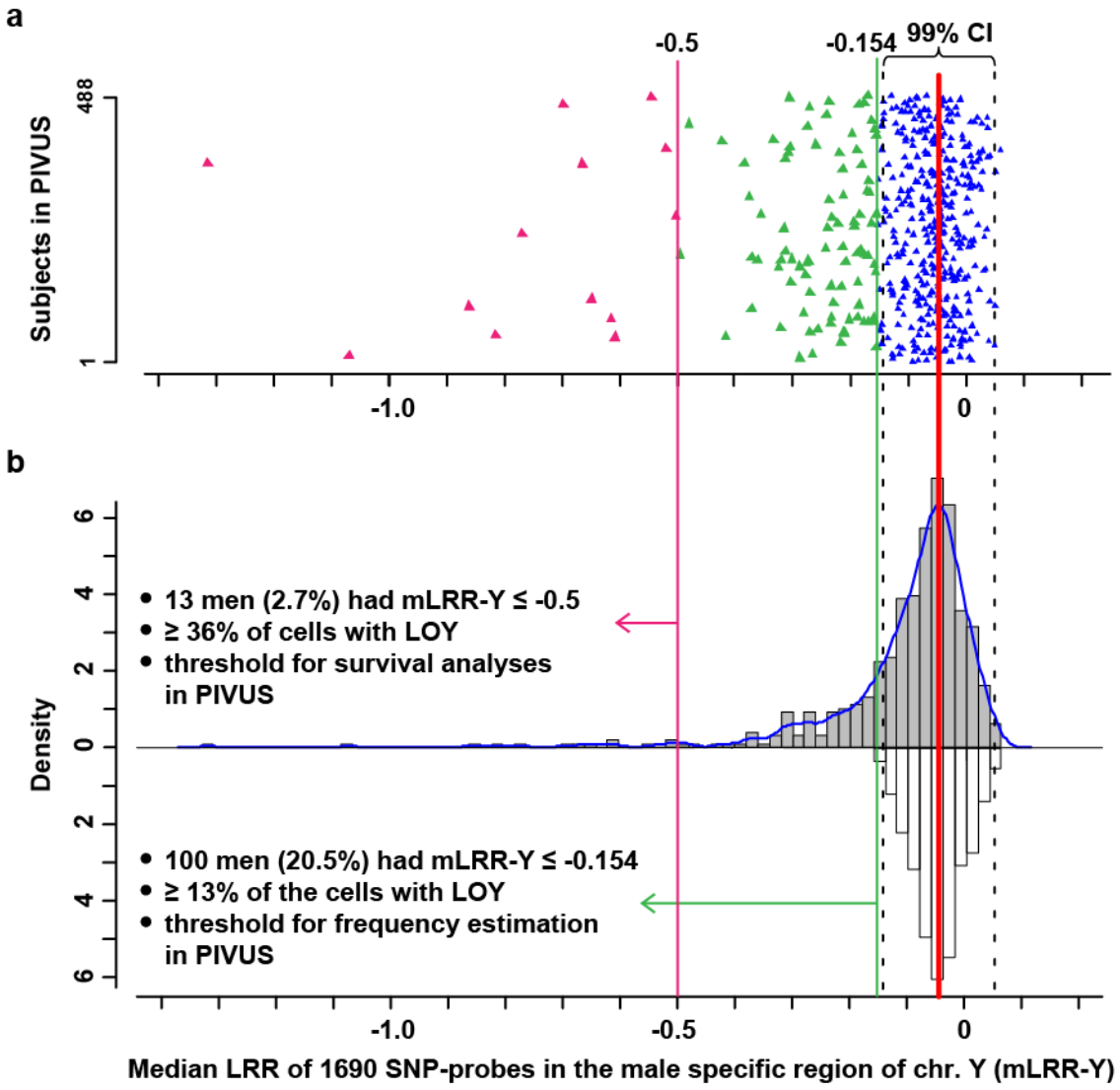
Supplementary Fig. 12



Supplementary Fig. 12. Results from exploratory survival analyses in the ULSAM cohort using Cox proportional hazards regression models with different thresholds for classification of participants into groups 1 and 0, based on their level of loss of chromosome Y (LOY) measured as the median Log R Ratio (LRR) in the male specific part of chromosome Y (mLRR-Y). The number of participants (n) with LOY and the minimum percentage of affected cells for each subject are given for each of the tested thresholds. The red and blue curves represent results from analyses with cancer mortality or all-cause mortality as endpoints, respectively. Models with significant effect on mortality (alpha level of 0.05) are indicated by solid black triangles and non-significant models are plotted with empty triangles. Based on these results, mLRR-Y at -0.4 is the most informative threshold for survival analyses in the studied ULSAM cohort.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 13

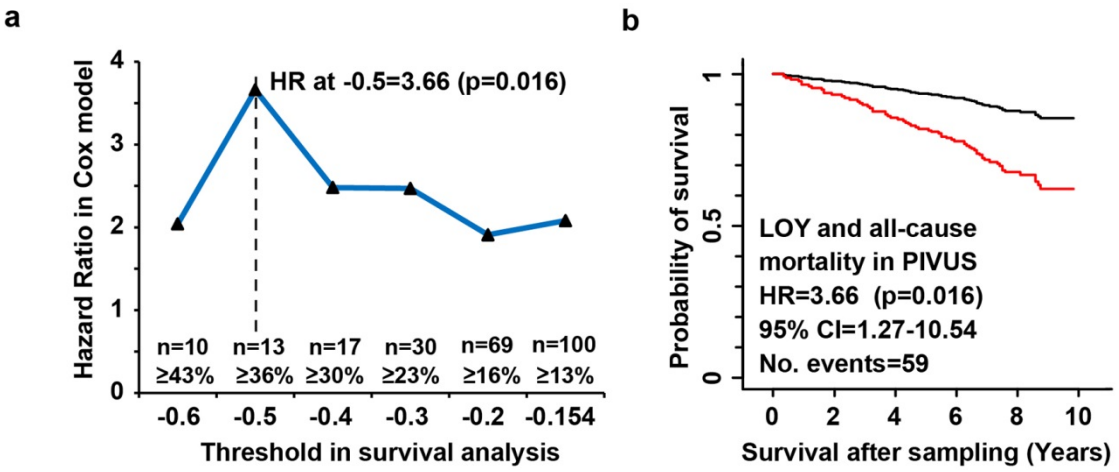


Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 13. LOY frequency estimation in PIVUS cohort after accounting for experimental variation. Panel a show the median Log R Ratio (LRR) in the male specific part of chromosome Y (mLRR-Y) observed in all men (n=488) genotyped from this cohort. Each triangle represents one participant. Panel b show the distribution of the mLRR-Y (grey bars) and the experimental noise (white bars) that were used to find the threshold for estimation of LOY frequency. The latter distribution was generated as described in online methods. The dotted black lines represent the 99% confidence intervals (CI) of the distribution of expected experimental background noise (white bars). Among the 488 men in PIVUS we found that 100 subjects (20.5%) had the mLRR-Y value lower than -0.154, which represent the lowest value in the distribution of experimental noise and corresponds to $\geq$13% LOY-cells.

Nature Genetics: doi:10.1038/ng.2966

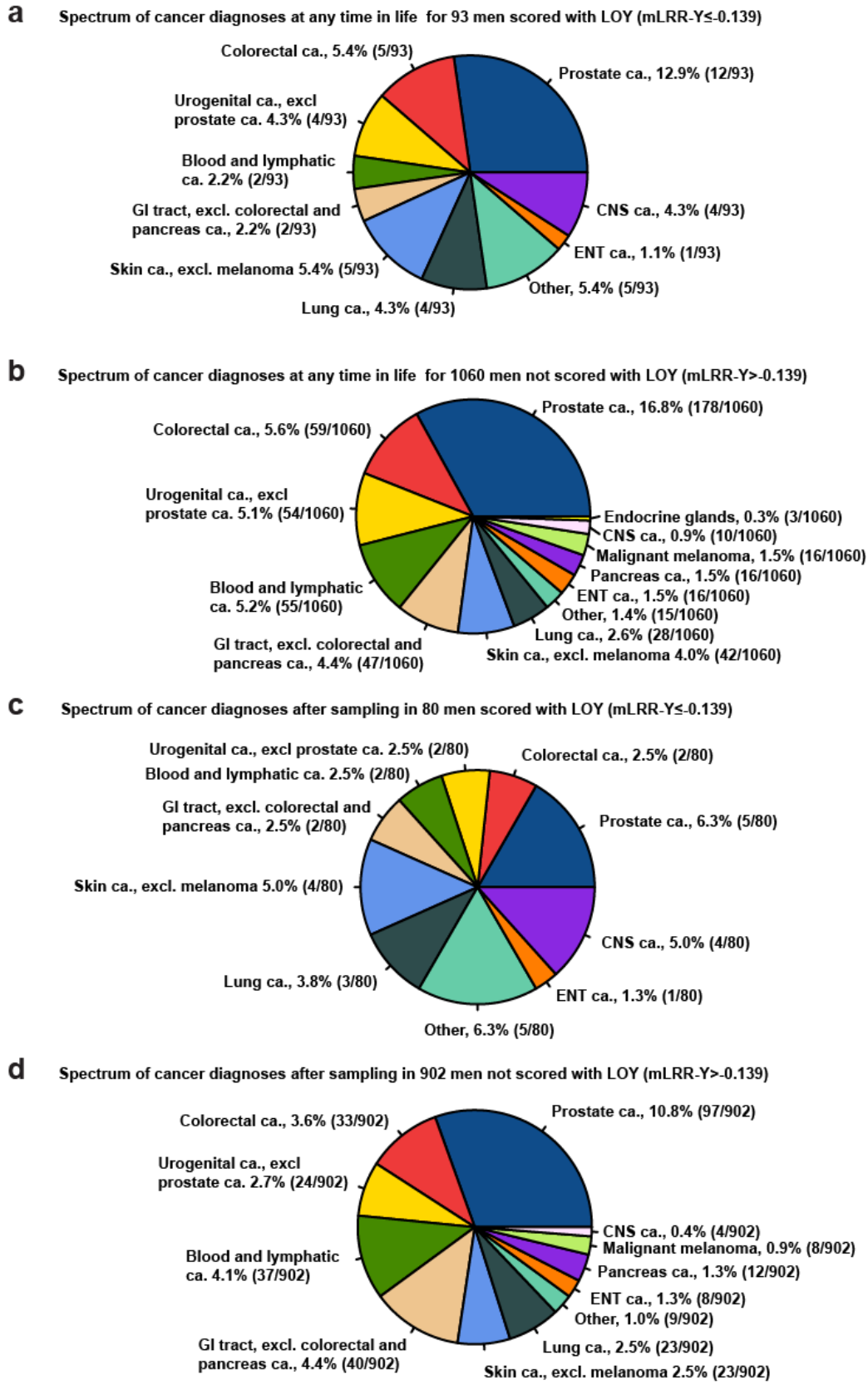
Supplementary Fig. 14



Supplementary Fig. 14. Validation of the result that men with loss of chromosome Y (LOY) are at a higher risk for all-cause mortality in an independent cohort (PIVUS). Panel a show results from Cox proportional hazards regression models with all-cause mortality as endpoint using different thresholds in analyses of 488 men. The participants were classified into groups 1 and 0 based on their degree of loss of chromosome Y (LOY) using different thresholds for mLRR-Y, i.e. the median Log R Ratio (LRR) in the male specific part of chromosome Y. The number of participants (n) with LOY and the minimum percentage of affected cells for each subject are given for each of the tested thresholds. Based on these results, mLRR-Y at -0.5 is the most informative threshold for survival analyses in the PIVUS cohort. Panel b shows results from a Cox proportional hazards regression model testing the effect from LOY on risk for all-cause mortality in 488 PIVUS men at the -0.5 threshold. The survival of men with LOY are represented in the red curve. Hazard ratio (HR), p-value, 95% confidence intervals (CI) and number of events and are shown.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 15



Nature Genetics: doi:10.1038/ng.2966

Supplementary Fig. 15. Comparisons of the spectrum of cancer diagnoses between ULSAM participants with and without LOY, which were successfully genotyped on Illumina beadchips and scored for structural genetic variants. The cut-off level for LOY used for these comparisons was mLRR-Y $\leq$-0.139, corresponding to LOY in $\geq$18% of cells (see **Fig. 2**, **Supplementary Fig. 3** and text). Panel a and b display cancer diagnoses for the entire cohort of 1153 men with 93 participants scored with LOY and 1060 subjects that were not scored with LOY, respectively. Panel c and d show the distribution of the cancer diagnoses in the cohort after excluding subjects with cancer before blood sampling. In the remaining 982 men, 80 subjects were scored with LOY and 902 subjects that were not scored with LOY. All cancer diagnoses were grouped into 13 categories, and for each category, the percentage of cases is shown, followed by the absolute number of patients (in parentheses) with this diagnosis category. GI – gastrointestinal; ENT – ear/nose/throat; CNS – central nervous system.

Nature Genetics: doi:10.1038/ng.2966

Supplementary Table 1. 40 autosomal structural aberrations ≥2 Mb detected in the ULSAM cohort and summarized in Figure 1.

Type	Chromosome	Coordinates *	Size (bp)	Figure
CNNLOH	1	chr1:0-51782000	51 782 000	Supplementary Fig. 10 f
CNNLOH	4	chr4:87780000-182786646	95 006 646	Supplementary Fig. 8 d
CNNLOH	6	chr6:0-26128000	26 128 000	
CNNLOH	7	chr7:0-0-159138663	159 138 663	
CNNLOH	9	chr9:99582000-141213431	41 631 431	Supplementary Fig. 8 c
CNNLOH	11	chr11:119974800-135006516	15 031 716	
CNNLOH	11	chr11:0-31285000	31 285 000	
CNNLOH	11	chr11:0-40329000	40 329 000	Supplementary Fig. 9 e
CNNLOH	14	chr14:24944467-107349540	82 405 073	
CNNLOH	14	chr14:40334000-107349540	67 015 540	Supplementary Fig. 9 f
CNNLOH	15	chr15:52081005-102520750	50 439 745	
CNNLOH	19	chr19:35480000-59128983	23 648 983	
CNNLOH	19	chr19:41725000-59128983	17 403 983	
CNNLOH	22	chr22:17000000-51304566	34 304 566	
CNNLOH	22	chr22:24700000-51304566	26 604 566	Supplementary Fig. 10 b
CNNLOH	22	chr22:26990000-51304566	24 314 566	Supplementary Fig. 10 d
Deletion	1	chr1:14922000-29839000	14 917 000	Supplementary Fig. 8 a
Deletion	2	chr2:206558932-209073561	2 514 629	
Deletion	4	chr4:1648000-35332000	33 680 000	Supplementary Fig. 8 e
Deletion	11	chr11:82565000-134950000	52 380 000	Supplementary Fig. 8 f
Deletion	13	chr13:34443925-52117369	17 673 444	
Deletion	13	chr13:41415834-51874437	10 458 603	Supplementary Fig. 8 b
Deletion	17	chr17:28036857-30282993	2 246 136	
Deletion	17	chr17:88964-21461735	21 372 771	
Deletion	20	chr20:30898371-44973847	14 075 476	
Deletion	20	chr20:31610411-43327038	11 716 627	Supplementary Fig. 9 b
Deletion	20	chr20:31785408-42077275	10 291 867	
Deletion	20	chr20:34543031-47267244	12 724 213	Supplementary Fig. 9 d
Deletion	20	chr20:43237779-47253475	4 015 696	Supplementary Fig. 10 a
Gain	3	chr3:139362907-142141603	2 778 696	
Gain	8	chr8:76206416-146166950	69 960 534	
Gain	9	chr9:0-141213431	141 213 431	Supplementary Fig. 10 c
Gain	9	chr9:70512737-141213431	70 700 694	
Gain	11	chr11:85044918-88006456	2 961 538	Supplementary Fig. 10 e
Gain	14	chr14:40485018-43137890	2 652 872	
Gain	15	chr15:20104200-102520750	82 416 550	Supplementary Fig. 9 c
Gain	15	chr15:20479640-23649047	3 169 407	
Gain	15	chr15:20104200-102520750	82 416 550	Supplementary Fig. 9 a
Gain	15	chr15:27659524-30378311	2 718 787	
Gain	15	chr15:64457910-102531392	38 073 482	

Notes:
CNNLOH - copy number neutral loss of heterozygozity
* positions according to hg19/GRCh37

Nature Genetics: doi:10.1038/ng.2966

Supplementary Table 2. Confounding factors for the 1153 analyzed ULSAM participants

	n	Median (SD)	% Y	% N	Proportion (%)
Age at sampling (Years)	1153	74.2 (3.5)			
Hypertension (Z102)	1109		74.0	26.0	
Exercise habits (Z106)	1010		96.5	3.5	
Smoking (X085)	1152		70.7	29.3	
Diabetes (Z378)	1109		10.9	89.1	
BMI (kg/m^2, Z290)	1106	25.9 (3.4)			
LDL (mmol/l, Z324)	1104	3.8 (0.9)			
HDL (mmol/l, Z302)	1108	1.2 (0.3)			
Education level (X098)	1112				
class 1					60.6
class 2					13.6
class 3					9.0
class 4					16.8

Notes: The confounding factors for the entire ULSAM cohort and the definitions of the variables used here can be found at http://www2.pubcare.uu.se/ULSAM/invest/indexinv.htm. The codes given after the variables in this table can be used to point to the correct information at the webpage. Data for age, BMI (body mass index), LDL (Low-Density Lipoprotein) and HDL (High-Density Lipoprotein) are given as medians with standard deviations (SD). Data describing hypertension, exercise habits, smoking and diabetes are given as proportions in columns marked "% Y" and "% N". Four classes of exercise habits were analyzed as N (sedentary=class 1, n=35) and Y (moderate, regular and athletic, classes 2-4, n=975). Three classes of smoking were analyzed as N (non-smoker=class 0, n=338) and Y (smokers and ex-smokers, classes 1 and 2, n=814).

Nature Genetics: doi:10.1038/ng.2966

Supplementary Table 3. 12 autosomal structural aberrations ≥2 Mb detected in the PIVUS cohort.

Type	Chromosome	Coordinates *	Size (bp)
CNNLOH	9	chr9:0-36251060	36 251 060
CNNLOH	12	chr12:51940000-133810000	81 870 000
CNNLOH	14	chr14:94156220-107331190	13 174 970
Deletion	2	chr2:24935579-27040368	2 104 789
Deletion	17	chr17:0-19238441	19 238 441
Deletion	20	chr20:31459525-45055839	13 596 314
Gain	12	chr12:0-133810000	133 810 000
Gain	12	chr12:0-133810000	133 810 000
Gain	16	chr16:15409870-18325001	2 915 131
Gain	2	chr2:152164587-154034610	1 870 023
Gain	3	chr3:80243013-83456039	3 213 026
Gain	5	chr5:0-10643429	10 643 429

Notes:
CNNLOH - copy number neutral loss of heterozygozity
* positions according to
hg19/GRCh37

Nature Genetics: doi:10.1038/ng.2966

Supplementary Table 4. Cox proportional hazards regression evaluating effect from LOY on all-cause mortality in 488 PIVUS men (no of events=59).

	HR	95% CI	P-value
Genotyping age	2.24	0.28-17.73	0.444
Hypertension	1.25	0.69-2.26	0.461
Exercise habits	0.71	0.49-1.01	0.057
Smoking	1.56	0.77-3.17	0.219
Diabetes	1.56	0.77-3.16	0.218
BMI	0.92	0.84-1.00	0.045 *
LDL-cholesterol	0.91	0.67-1.24	0.555
HDL-cholesterol	0.78	0.37-1.66	0.527
Education level	0.91	0.67-1.23	0.534
Autosomal LOH (>2 Mb)	2.66	0.05-2.85	0.344
LOY	5.24	1.27-21.58	0.022 *

Notes: HR – hazard ratio. 95% CI – 95% confidence interval. Autosomal LOH (>2 Mb) – autosomal loss of heterozygozity; this category was composed of deletions and CNNLOH events larger than 2 million bp. The effect of autosomal gains could not be estimated as in the ULSAM cohort because none of the subject with autosomal gains >2Mb had died during follow up time. The longest and the median follow-up time was >10 and 7.0 years, respectively. DNA extracted from blood of 488 men (median age 70) were genotyped using Illumina Omni-Express chip (containing ~730 000 SNP-probes). A continuous explanatory variable was used as a proxy for loss of chromosome Y (mLRR-Y, see text and **Supplementary Fig. 13**). * - indicates statistically significant effects with 0.05 alpha value.

Nature Genetics: doi:10.1038/ng.2966